CLAIMS

What is claimed is:

1. A method of determining a sequence of a nucleic acid polymer, comprising the steps of:

(a) obtaining data traces from a plurality of channels of an electrophoresis detection apparatus, each channel detecting the products of a DNA sequencing reactions;

(b) combining the data traces by a process comprising the steps of:

(i) applying a cross-correlation coefficient to each of the four data traces to yield four refined traces, wherein the cross-correlation coefficient compares each of the traces with an ideal, Gaussian-shaped peak and wherein the refined traces have narrower peaks than the corresponding data traces;

(ii) combining the four refined traces to yield a composite trace;

(c) detecting peaks in the composite trace by a process that is independent of peak spacing; and

(d) determining the sequence of the nucleic acid polymer by assigning basecalls to the peaks.

2. The method of claim 1, wherein the data traces have been preprocessed.

3. The method of claim 2, wherein preprocessing comprises the steps of:

(i) obtaining unprocessed data traces from a plurality of channels of an automated electrophoresis detection apparatus, each channel detecting the products of a DNA sequencing reactions;

(ii) identifying begin and end points in the unprocessed data;

(iii) establishing a baseline in the unprocessed data,

(iv) subtracting the baseline from the unprocessed data to generate the baseline-corrected data; and

(v) separating the baseline-corrected data to generate the data traces, the separating step comprising spectral or leakage separation.

4. The method of claim 1, wherein the electrophoresis detection apparatus uses slab gel electrophoresis; tube gel electrophoresis; or capillary gel electrophoresis.

5. The method of claim 4, wherein the electrophoresis detection apparatus is a MegaBACE capillary sequencing machine.

6. The method of claim 1, further comprising the step of generating at least one quality score for at least one basecall.

7. The method of claim 6, wherein the at least one quality score is a gap-quality score, wherein the gap-quality score estimates the probability of a deletion error between two adjacent assigned basecalls.

8. The method of claim 7, wherein the gap-quality score measures degree of noise between the two adjacent assigned basecalls and overly wide peak spacing between the two adjacent assigned basecalls.

9. The method of claim 6, further comprising the step of using the quality scores for quality filtering whereby basecalls can be removed or added from the sequence of the nucleic acid polymer during the quality filtering.

10. The method of claim 1, wherein the DNA sequencing reactions utilize dye-terminator or dye-primer chemistry.

11. A computer program product comprising a machine readable medium on which is provided program instructions for determining a sequence of a nucleic acid polymer, the instructions comprising:

code for obtaining data traces from a plurality of channels of an electrophoresis detection apparatus, each channel detecting the products of a DNA sequencing reactions;

code for combining the data traces by a process comprising the steps of:

(i) applying a cross-correlation coefficient to each of the four data traces to yield four refined traces, wherein the cross-correlation coefficient compares each of the traces with an ideal, Gaussian-shaped peak and wherein the refined traces have narrower peaks than the corresponding data traces;

(ii) combining the four refined traces to yield a composite trace;

code for detecting peaks in the composite trace by a process that is independent of peak spacing; and

code for determining the sequence of the nucleic acid polymer by assigning basecalls to the peaks.

12. The computer program product of claim 11, wherein the data traces have been preprocessed.

13. The computer program product of claim 12, wherein preprocessing comprises the steps of:

(i) obtaining unprocessed data traces from a plurality of channels of an automated electrophoresis detection apparatus, each channel detecting the products of a DNA sequencing reactions;

(ii) identifying begin and end points in the unprocessed data;

(iii) establishing a baseline in the unprocessed data,

(iv) subtracting the baseline from the unprocessed data to generate the baseline-corrected data; and

(v) separating the baseline-corrected data to generate the data traces, the separating step comprising spectral or leakage separation.

14. The computer program product of claim 11, wherein the electrophoresis detection apparatus uses slab gel electrophoresis; tube gel electrophoresis; or capillary gel electrophoresis.

15. The computer program product of claim 14, wherein the electrophoresis detection apparatus is a MegaBACE capillary sequencing machine.

16. The computer program product of claim 11, further comprising code for generating at least one quality score for at least one basecall.

17. The computer program product of claim 16 wherein the at least one quality score is a gap-quality score, wherein the gap-quality score estimates the probability of a deletion error between two adjacent assigned basecalls.

18. The computer program product of claim 17, wherein the gap-quality score measures degree of noise between the two adjacent assigned basecalls and overly wide peak spacing between the two adjacent assigned basecalls.

19. The computer program product of claim 16, further comprising code for using the quality scores for quality filtering whereby basecalls can be removed or added from the sequence of the nucleic acid polymer during the quality filtering.

20. A computing device comprising a memory device configured to store at least temporarily program instructions for determining a sequence of a nucleic acid polymer, the instructions comprising:

code for obtaining data traces from a plurality of channels of an electrophoresis detection apparatus, each channel detecting the products of a DNA sequencing reactions;

code for combining the data traces by a process comprising the steps of:

(i) applying a cross-correlation coefficient to each of the four data traces to yield four refined traces, wherein the cross-correlation coefficient compares each of the traces with an ideal, Gaussian-shaped peak and wherein the refined traces have narrower peaks than the corresponding data traces;

(ii) combining the four refined traces to yield a composite trace;

code for detecting peaks in the composite trace by a process that is independent of peak spacing; and

code for determining the sequence of the nucleic acid polymer by assigning basecalls to the peaks.

21. The computing device of claim 20, wherein the data traces have been preprocessed.

22. The computing device of claim 21, wherein preprocessing comprises the steps of:

(i) obtaining unprocessed data traces from a plurality of channels of an automated electrophoresis detection apparatus, each channel detecting the products of a DNA sequencing reactions;

(ii) identifying begin and end points in the unprocessed data;

(iii) establishing a baseline in the unprocessed data,

(iv) subtracting the baseline from the unprocessed data to generate the baseline-corrected data; and

(v) separating the baseline-corrected data to generate the data traces, the separating step comprising spectral or leakage separation.

23. The computing device of claim 20, wherein the electrophoresis detection

apparatus uses slab gel electrophoresis; tube gel electrophoresis; or capillary gel electrophoresis.

24. The computing device of claim 23, wherein the electrophoreis detection apparatus is a MegaBACE capillary sequencing machine.

25. The computing device of claim 20, further comprising code for generating at least one quality score for at least one basecall.

26. The computing device of claim 25, wherein the at least one quality score is a gap-quality score, wherein the gap-quality score estimates the probability of a deletion error between two adjacent assigned basecalls.

27. The computing device of claim 26, wherein the gap-quality score measures degree of noise between the two adjacent assigned basecalls and overly wide peak spacing between the two adjacent assigned basecalls.

28. The computing device of claim 25, further comprising code for using the quality scores for quality filtering whereby basecalls can be removed or added from the sequence of the nucleic acid polymer during the quality filtering.

29. A method for estimating the probability that a basecall was missed between two adjacent assigned basecalls, the method comprising the steps of:

(a) measuring degree of noise between the two adjacent assigned basecalls;

(b) measuring peak spacing between the two adjacent assigned basecalls; and

(c) computing a gap-quality score, wherein the gap-quality score estimates the probability that a base call was missed between the two adjacent assigned basecalls.

30. The method of claim 29, wherein the basecalls were assigned using the method of claim 1.

31. A computer program product comprising a machine readable medium on which is provided program instructions for estimating the probability that a basecall was missed between two adjacent assigned basecalls, the instructions comprising:

code for measuring degree of noise between the two adjacent assigned basecalls;

code for measuring peak spacing between the two adjacent assigned basecalls; and

code for computing a gap-quality score, wherein the gap-quality score estimates the probability that a base call was missed between the two adjacent assigned basecalls.

32. A computing device comprising a memory device configured to store at least temporarily program instructions estimating the probability that a basecall was missed between two adjacent assigned basecalls, the instructions comprising:

code for measuring degree of noise between the two adjacent assigned basecalls;

code for measuring peak spacing between the two adjacent assigned basecalls; and

code for computing a gap-quality score, wherein the gap-quality score estimates the probability that a base call was missed between the two adjacent assigned basecalls.

33. A method for benchmarking basecaller performance, the method comprising the steps of:

(a) determining a nucleic acid sequence using two basecalling algorithms to yield two test sequences;

(b) identifying an aligned sequence between the two test sequences using a sequence comparison algorithm;

(c) comparing the sequence of each of the test sequences with the aligned sequence using the sequence comparison algorithm;

(d) determining high quality left-most and right-most alignments from the comparison;

(e) extending the aligned sequence by identifying a left-most and right-most boundary wherein such boundaries correspond to the left-most and right-most alignments, respectively; and

(f) collecting error statistics over the extended aligned sequence between its left-most and right-most boundaries.

34. The method of claim 33, wherein the sequence comparison algorithm is Blast.

35. The method of claim 33, wherein the error statistics are derived in conjunction with quality scores.

36. The method of claim 35, wherein the quality scores are call quality scores and gap-quality scores.

37. The method of claim 36, wherein preference is given to high call quality scores.

38. The method of claim 36, wherein a low gap-quality indicates a high probability for a deletion error.

39. The method of claim 36, wherein substitution errors are linked to the call

quality scores.

40. The method of claim 36, wherein insertion errors are linked to the call quality scores.

41. A computer program product comprising a machine readable medium on which is provided program instructions for benchmarking basecaller performance, the instructions comprising:

code for determining a nucleic acid sequence using two basecalling algorithms to yield two test sequences;

code for identifying an aligned sequence between the two test sequences using a sequence comparison algorithm;

code for comparing the sequence of each of the test sequences with the aligned sequence using the sequence comparison algorithm;

code for determining high quality left-most and right-most alignments from the comparison;

code for extending the aligned sequence by identifying a left-most and right-most boundary wherein such boundaries correspond to the left-most and right-most alignments, respectively; and

code for collecting error statistics over the extended aligned sequence between its left-most and right-most boundaries.

42. The computer program product of claim 41, wherein the sequence comparison algorithm is Blast.

43. The computer program product of claim 41, wherein the error statistics are derived in conjunction with quality scores.

44. The computer program product of claim 43, wherein the quality scores are call

quality scores and gap-quality scores.

45. The computer program product of claim 44, wherein preference is given to high call quality scores.

46. The computer program product of claim 43, wherein a low gap-quality indicates a high probability for a deletion error.

47. The computer program product of claim 44, wherein substitution errors are linked to the call quality scores.

48. The computer program product of claim 44, wherein insertion errors are linked to the call quality scores.

49. A computing device comprising a memory device configured to store at least temporarily program instructions for benchmarking basecaller performance, the instructions comprising:

code for determining a nucleic acid sequence using two basecalling algorithms to yield two test sequences;

code for identifying an aligned sequence between the two test sequences using a sequence comparison algorithm;

code for comparing the sequence of each of the test sequences with the aligned sequence using the sequence comparison algorithm;

code for determining high quality left-most and right-most alignments from the comparison;

code for extending the aligned sequence by identifying a left-most and right-most boundary wherein such boundaries correspond to the left-most and right-most alignments, respectively; and

code for collecting error statistics over the extended aligned sequence between its left-most and right-most boundaries.

50. The computing device of claim 49, wherein the sequence comparison algorithm is Blast.

51. The computing device of claim 49, wherein the error statistics are derived in conjunction with quality scores.

52. The computing device of claim 51, wherein the quality scores are call quality scores and gap-quality scores.

53. The computing device of claim 52, wherein preference is given to high call quality scores.

54. The computing device of claim 52, wherein a low gap-quality indicates a high probability for a deletion error.

55. The computing device of claim 52, wherein substitution errors are linked to the call quality scores.

56. The computing device of claim 52, wherein insertion errors are linked to the call quality scores.

57. A call quality score, said score being estimated through a process that relies on continuously varying parameters of basecall quality.

58. The call quality score of Claim 57, wherein said process does not utilize a look-up table.

59. A gap-quality score, said score estimating the probability that a basecall was missed following a given assigned basecall.

60. The gap-quality score of Claim 59, wherein said assigned basecall is determined by the method of Claim 1.

61. The gap-quality score of Claim 59, wherein said score is derived from the method of Claim 29.